

Terpene metabolic engineering *via* nuclear or chloroplast genomes profoundly and globally impacts off-target pathways through metabolite signalling

Elise K. Pasoreck¹, Jin Su¹, Ian M. Silverman², Sager J. Gosai², Brian D. Gregory², Joshua S. Yuan³ and Henry Daniell^{1*}

¹Department of Biochemistry, School of Dental Medicine, University of Pennsylvania, Philadelphia, PA, USA

²Department of Biology, University of Pennsylvania, Philadelphia, PA, USA

³Department of Plant Pathology and Microbiology, Texas A&M University, College Station, TX, USA

Received 11 November 2015;

revised 3 February 2016;

accepted 8 February 2016.

*Correspondence (Tel 215 746 2563; fax 215 898 3695; email hdaniell@upenn.edu)

GEO accession number for RNA-seq data: GSE74103.

Keywords: Secondary metabolism, retrograde/anterograde signalling, transgenic/transplastomic plants, squalene.

Summary

The impact of metabolic engineering on nontarget pathways and outcomes of metabolic engineering from different genomes are poorly understood questions. Therefore, squalene biosynthesis genes *FARNESYL DIPHOSPHATE SYNTHASE* (*FPS*) and *SQUALENE SYNTHASE* (*SQS*) were engineered *via* the *Nicotiana tabacum* chloroplast (C), nuclear (N) or both (CN) genomes to promote squalene biosynthesis. *SQS* levels were ~4300-fold higher in C and CN lines than in N, but all accumulated ~150-fold higher squalene due to substrate or storage limitations. Abnormal leaf and flower phenotypes, including lower pollen production and reduced fertility, were observed regardless of the compartment or level of transgene expression. Substantial changes in metabolomes of all lines were observed: levels of 65–120 unrelated metabolites, including the toxic alkaloid nicotine, changed by as much as 32-fold. Profound effects of transgenesis on nontarget gene expression included changes in the abundance of 19 076 transcripts by up to 2000-fold in CN; 7784 transcripts by up to 1400-fold in N; and 5224 transcripts by as much as 2200-fold in C. Transporter-related transcripts were induced, and cell cycle-associated transcripts were disproportionately repressed in all three lines. Transcriptome changes were validated by qRT-PCR. The mechanism underlying these large changes likely involves metabolite-mediated anterograde and/or retrograde signalling irrespective of the level of transgene expression or end product, due to imbalance of metabolic pools, offering new insight into both anticipated and unanticipated consequences of metabolic engineering.

Introduction

Many metabolic engineering strategies rely on transformation of the nuclear genome. For example, avian *FARNESYL DIPHOSPHATE SYNTHASE* (*FPS*) and yeast *SQUALENE SYNTHASE* (*SQS*) have been expressed in *Nicotiana tabacum* to promote squalene accumulation (Wu *et al.*, 2012). New metabolic pathways have also been engineered, for example by introducing the pathway for the cyanogenic glucoside dhurrin into *Arabidopsis* (Tattersall *et al.*, 2001). Several other studies have found that directing key enzymes to chloroplasts can substantially increase the content of metabolites such as β -carotene—used to produce ‘golden rice’ (Ye *et al.*, 2000)—and terpenes, including patchoulol (Wu *et al.*, 2006) and squalene (Wu *et al.*, 2012).

However, nuclear transformation is associated with several disadvantages. Positional effects can reduce or silence transgene expression, making it difficult to establish a stable population that consistently accumulates high levels of the desired product (Jin and Daniell, 2015; Verma and Daniell, 2007). Additionally, lack of transgene containment raises regulatory issues and makes commercialization more challenging (Chan and Daniell, 2015). Some of these problems can be overcome by expressing transgenes via

the chloroplast genome. Chloroplast transformation, which occurs through homologous recombination, contains transgenes (Verma *et al.*, 2008), and its high ploidy—up to 10 000 copies per cell—permits transgene products to reach exceptionally high levels (Ruhlman *et al.*, 2010). For example, tobacco chloroplasts expressing a multigene operon encoding Bt toxin allowed the protein to accumulate up to ~50% of total leaf protein and to form cuboidal crystals within chloroplasts (De Cosa *et al.*, 2001); chloroplast-expressed β -glucosidase accumulated to 160-fold higher levels than in untransformed plants (Jin *et al.*, 2011); and proinsulin could make up nearly 70% of total transplastomic leaf protein (Ruhlman *et al.*, 2010).

Chloroplast genome transformation has also been successfully used in metabolic engineering (Jin and Daniell, 2015): chloroplasts expressing chorismate pyruvate-lyase accumulated the biopolymer *p*-hydroxybenzoic acid to ~26.5% dry weight—the highest level reported for any bioproduct—without compromising plant health (Viitanen *et al.*, 2004). In later studies, the chloroplast genome was engineered to express all six genes of the cytoplasmic mevalonate pathway for terpenoid synthesis (Kumar *et al.*, 2012) or thirteen artemisinin pathway biosynthesis genes (Saxena *et al.*, 2014), although in the latter case, expression was

Please cite this article as: Pasoreck, E.K., Su, J., Silverman, I.M., Gosai, S.J., Gregory, B.D., Yuan, J.S. and Daniell, H. (2016) Terpene metabolic engineering *via* nuclear or chloroplast genomes profoundly and globally impacts off-target pathways through metabolite signalling. *Plant Biotechnol. J.*, doi: 10.1111/pbi.12548

inadequate to meet target levels. However, chloroplast genome transformation requires further optimization, and a comparison of metabolic engineering outcomes *via* engineering different cellular compartments has not been performed.

The chloroplast genome is highly reduced, with many genes lost or transferred to the nucleus (Jensen and Leister, 2014). Consequently, chloroplast function requires the import of thousands of nuclear-encoded proteins, many of which work in concert with plastid-encoded gene products and require proper stoichiometry (Jin and Daniell, 2015). Therefore, the expression of nuclear- and plastid-encoded genes must be coordinately regulated, and this occurs *via* anterograde signalling from the nucleus to plastids and retrograde signalling from plastids to the nucleus. Whereas anterograde signalling is well understood, retrograde signalling is still enigmatic. Chloroplasts may regulate nuclear gene expression *via* proteins (Jin and Daniell, 2014; Singh *et al.*, 2008), redox state (Nott *et al.*, 2006) or metabolites (Chi *et al.*, 2015; Estavillo *et al.*, 2011; Woodson *et al.*, 2011; Xiao *et al.*, 2012), but these signals act only under specific circumstances, and mechanisms by which they are conveyed remain elusive. However, high accumulation of proteins expressed *via* the chloroplast genome and compartmentalization within chloroplasts make chloroplast genetic engineering an excellent system to study retrograde signalling.

Much work on metabolic engineering and synthetic biology has focused on engineering pathways to generate high-value metabolites, but the global impact of such engineering has not yet been explored despite the potential for unintended consequences (Bobik and Burch-Smith, 2015). The availability of modern tools to study the metabolome and transcriptome facilitates global evaluation of the effect of these introduced pathways on native genes through metabolite-mediated anterograde or retrograde signalling. Here, we used chloroplast genetic engineering and an existing nuclear transgenic line to uncover potential unintended consequences of expressing metabolic genes from different compartments. We focused on squalene because of its importance in steroid biosynthesis, its industrial applications in cosmetics and nutraceuticals (Kim and Karadeniz, 2012) and its use as a vaccine adjuvant (O'Hagan *et al.*, 2011), as well as the availability of nuclear transgenic lines expressing squalene biosynthetic enzymes for comparative investigations (Wu *et al.*, 2012). We found profound but similar off-target effects on both the metabolome and transcriptome regardless of the compartment from which transgenes were expressed. By focusing on global effects of intercompartmental signalling rather than attempting to identify yet another signalling molecule, we provide a framework for future studies on large-scale effects of metabolite-mediated intercompartmental signalling.

Results

Characterization of transplastomic lines expressing *FPS* and *SQS*

To engineer squalene biosynthesis through different organellar genomes, we generated two tobacco chloroplast expression vectors, one encoding Flag-tagged SQS (pLD-SQS) and one encoding both Flag-tagged SQS and His-tagged FPS (pLD-FPS-SQS) (Figure 1a). Encoded amino acid sequences were identical to those reported previously (Wu *et al.*, 2012), but corresponding DNA sequences were codon-optimized for enhanced chloroplast expression (Daniell *et al.*, 2009). In pLD-SQS, *Flag-SQS* was regulated by the tobacco *psbA* promoter, 5'-UTR and 3'-UTR,

and isoleucine tRNA (*trnI*) and alanine tRNA (*trnA*) flanking sequences were included for integration into the chloroplast genome *via* homologous recombination (Verma and Daniell, 2007). In pLD-FPS-SQS, expression of *His-FPS* was controlled by the plastid rRNA operon promoter (*Prrn*), the 5' translation control element of bacteriophage T7 gene 10 and the tobacco *rbcL* 3'-UTR (Dhingra *et al.*, 2004).

Two *N. tabacum* lines were used: the wild-type (WT) 1068 introduction, which has abundant glandular trichomes that may be squalene sinks (Wu *et al.*, 2012), and a transgenic line in the WT background expressing chloroplast-targeted SQS and FPS *via* the nuclear genome (Wu *et al.*, 2012), hereafter referred to as 'N'. Leaves of N were bombarded with pLD-SQS to generate transplastomic lines expressing SQS from the chloroplast genome; the resulting line is referred to as 'CN'. After selection and regeneration on spectinomycin-containing media, several independent CN lines were regenerated from ten bombardments. PCR analysis with the 3P/3M primer set (Verma *et al.*, 2008) indicated that the *Flag-SQS* cassette had been stably integrated into the chloroplast genome *via* homologous recombination (Figure S1). After two additional rounds of selection on spectinomycin-containing media, CN lines were confirmed by Southern blot. As shown in Figure 1a,b, several independent CN lines showed a 6.63-kb band but not the 4.43-kb WT band, confirming that homoplasmic plants had been generated. We also bombarded N with pLD-FPS-SQS, but no shoots survived. When pLD-FPS-SQS was used to transform WT tobacco, several independent transplastomic lines, referred to as 'C', were obtained, as confirmed by PCR (Figure S1) and Southern blot (Figure 1b, left). Two bands with correct sizes of 8.42 and 3.48 kb (due to the presence of a *Hind*III site in the *His-FPS* cassette), but not the 7.67-kb WT band, were observed in *Hind*III-digested DNA from C lines, confirming that homoplasmic C lines had been generated (Figure 1b, right). Because the engineered His tag was not detectable in western blots, we confirmed expression of *His-FPS* *via* the chloroplast genome using northern blot for the *rbcL* 3'-UTR (Figure 1a). In addition to the endogenous *rbcL* signal in each line, the C line showed an additional band corresponding to the expressed *His-FPS* transgene (Figure 1c). Notably, the *His-FPS* transcript was as abundant as that of *rbcL*, which is the most abundant protein on earth (Dhingra *et al.*, 2004).

Expression of Flag-SQS protein was detected by western blot using a monoclonal anti-Flag antibody. A Flag signal could be detected in both transgenic and transplastomic lines. However, detection of the signal in N required loading as much as 120-fold more protein (Figure 1d). To quantify this difference, we performed densitometric analysis. After accounting for differences in the amount of loaded protein and normalizing to N, we found that CN expressed 2813–4372 times more Flag-SQS than N and that C expressed 1399–2309 times more than N.

The impact of *FPS* and *SQS* expression on leaf and flower development

Regardless of the genome from which transgenes were expressed or levels of Flag-SQS, expression of *FPS* and *SQS* had a profound effect on leaf and flower development, but CN displayed most severe leaf and flower phenotypes. At the time of transfer to soil, CN leaves were half as long as WT (Figure 2a) and remained small, both after transferring to the glasshouse (Figure 2b) and at the onset of flowering (Figure 2c). In particular, leaves of adult CN plants were shorter, narrower and more curled (Figure 2c). Leaves of N were also initially shorter than WT (Figure 2a,b), but

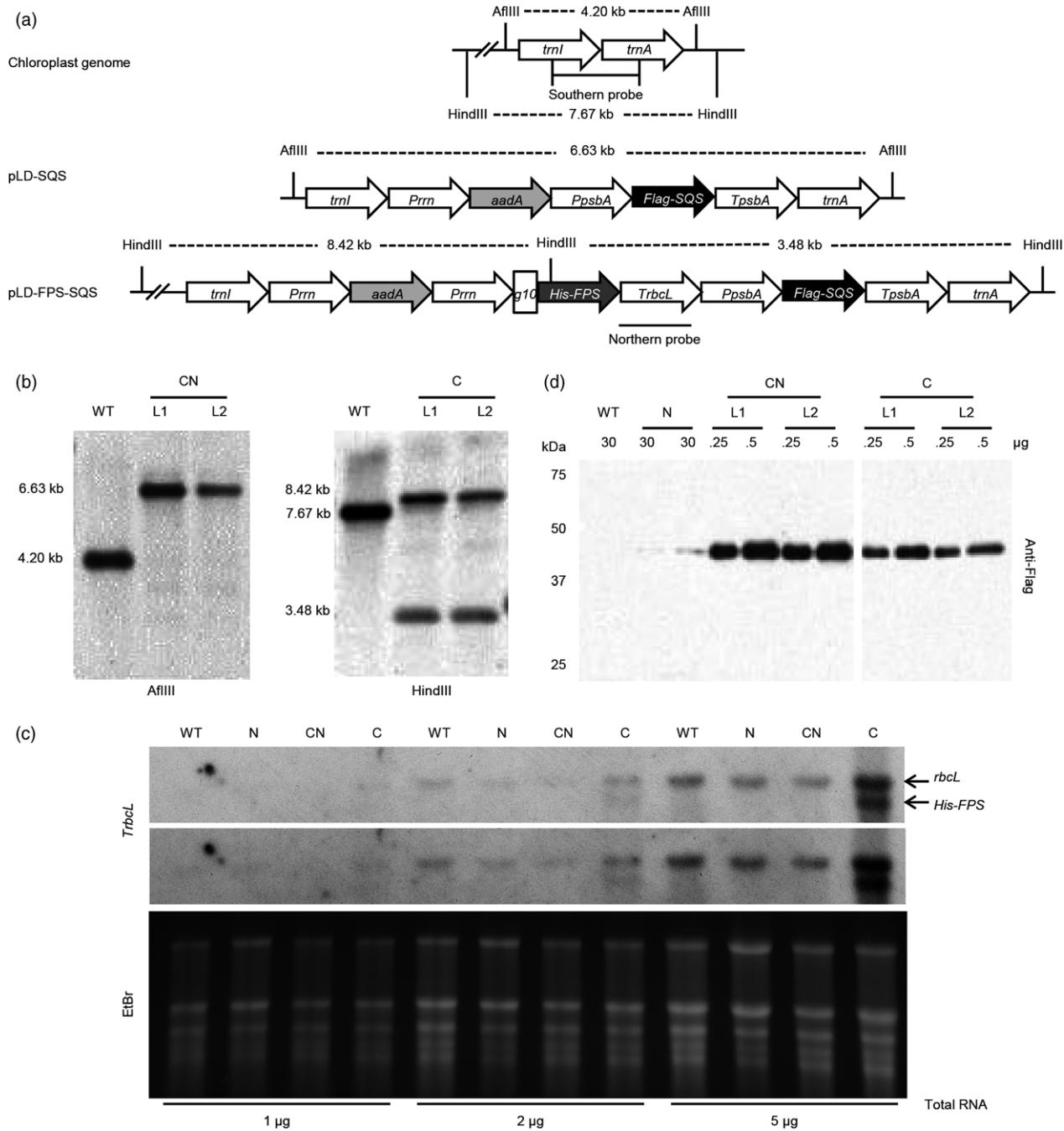


Figure 1 Chloroplast transformation vectors and characterization of transplastomic lines. (a) Schematics of the tobacco chloroplast genome, transformation vectors and probe hybridization sites with predicted sizes of hybridizing fragments in Southern blots. Top, untransformed; centre, pLD-SQS; bottom, pLD-FPS-SQS. (b) Genomic DNA (0.5 µg) from independent transplastomic lines was digested with *Af*III (CN lines, left) or *Hind*III (C lines, right) and probed with a DIG-labelled probe specific for the *trnL/trnA* flanking region. (c) Total RNA (1–5 µg) was hybridized with a DIG-labelled *TrbC1* probe. Northern blots were exposed to films for 15 min. (top) or 1 h. (middle), and ethidium bromide-stained gel (bottom) is a loading control. (d) Anti-Flag western blot of indicated amounts of total protein. FPS, FARNESYL DIPHOSPHATE SYNTHASE; SQS, SQUALENE SYNTHASE.

as they aged, leaf morphology more closely resembled that of WT (Figure 2c). A similar phenomenon was observed for C plants, whose pleiotropic phenotypes became less pronounced with age (Figure 2b,c).

Inflorescences of engineered lines were more clustered than those of WT (Figure 2d), and many buds aborted prematurely (Figure 2e); these buds contained tiny, underdeveloped stigmas and desiccated petals and stamens (Figure 2e). In CN flowers,

stamens were on long, curved filaments that produced minimal pollen (Figure 2f). By contrast, N and C flowers displayed stamens on straight filaments with anthers that contained noticeably more pollen (Figure 2g,h), although N filaments were generally longer and anthers contained more pollen than C (compare Figure 2g with Figure 2h). Unlike in engineered lines, WT stamens were the same length as stigmas, and anthers produced abundant pollen (Figure 2i).

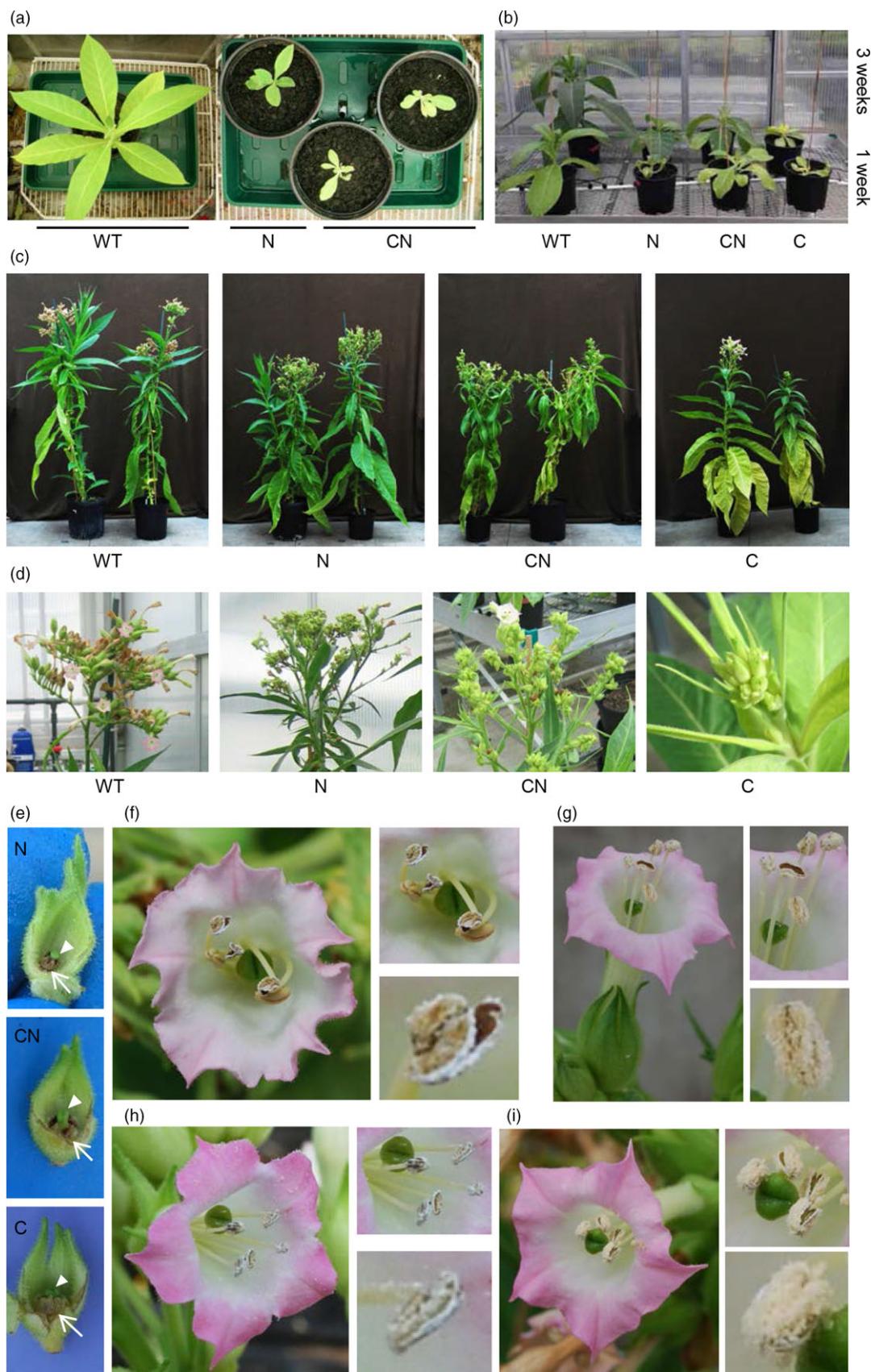


Figure 2 Leaf and flower phenotypes of FPS/SQS plants. (a) Plants after 5 weeks in the growth chamber. (b, c) Glasshouse-grown plants after 1 week (b, front), 3 weeks (b, back) or 5 months (c). (d) Shoot apices. (e) Aborted buds. (f-i) Phenotypes of CN (f), N (g), C (h) and WT (i) flowers. FPS, FARNESYL DIPHOSPHATE SYNTHASE; SQS, SQUALENE SYNTHASE; WT, wild-type

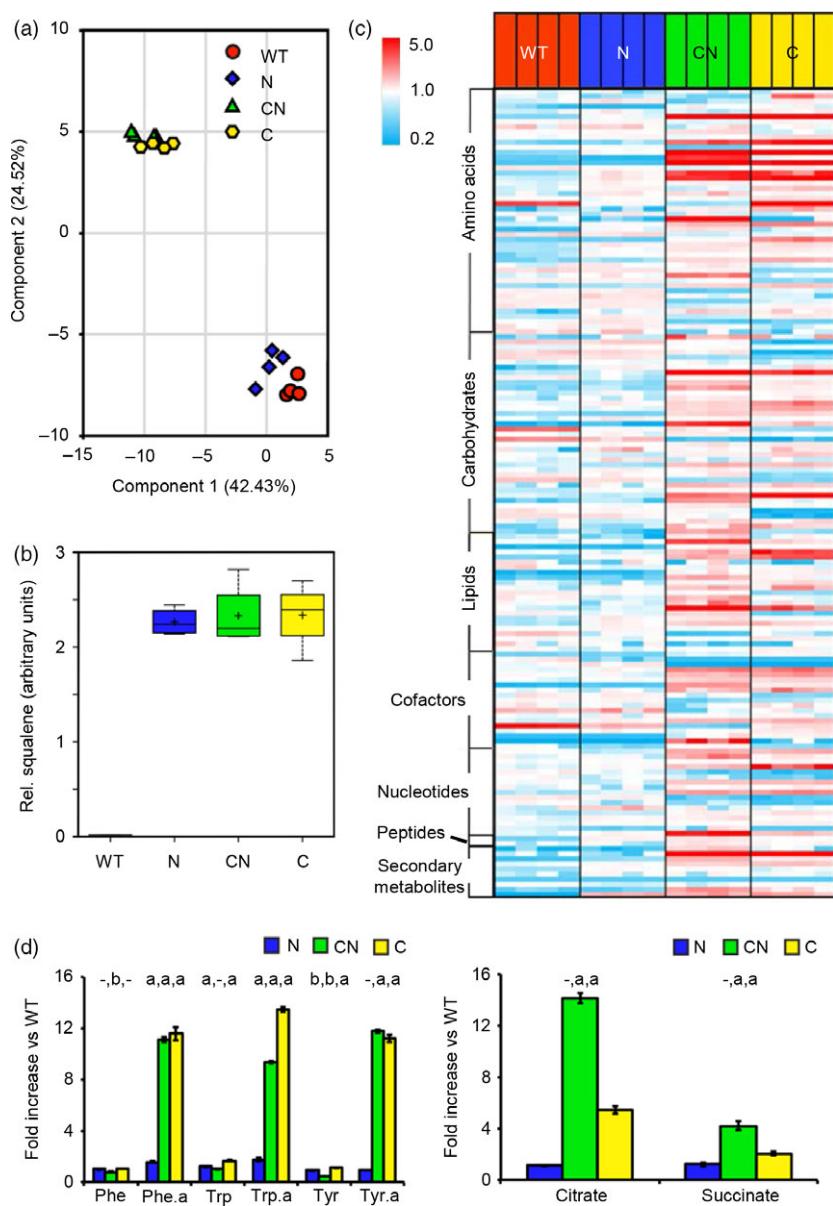
Notably, because of differences in antibiotic selection when plants were initially grown (WT and N on antibiotic-free media and CN and C on spectinomycin-containing media), it is more meaningful to compare CN with C and WT with N. Given this, differences between CN and C are quite striking; despite both expressing high levels of Flag-SQS, CN had much more severe phenotypes. Similarly, N's phenotypes were as clear as those of CN and C and occurred despite expressing comparatively little Flag-SQS.

Evaluation of metabolic profiles of transgenic/transplastomic lines expressing *FPS* and *SQS*

Although our original intention was to produce transplastomic plants that accumulated squalene, phenotypes of all three lines suggested that more than just squalene levels might be altered. Therefore, we used nodal cutting to grow all lines *in vitro* under identical physiological conditions and harvested the third youngest leaf from plants at an identical developmental stage to analyse global metabolic profiles. All transplastomic plants were from the

T0 generation. The metabolomic analysis used gas chromatography/mass spectrometry (GC/MS) and two types of ultra high-performance liquid chromatography/tandem mass spectrometry (UHPLC/MS/MS), one optimized for acidic species and one optimized for basic species (Clarke *et al.*, 2013; Evans *et al.*, 2009). Principal component analysis (PCA) of quadruplicate samples showed tight clustering within each line (Figure 3a). PCA also showed that metabolic profiles of CN and C were distinct from N and WT but were almost indistinguishable from each other, whereas N showed a small but clear separation from WT (Figure 3a).

Metabolomic-based analysis of squalene levels indicated that despite up to 4000-fold higher Flag-SQS levels in CN and C than in N (Figure 1d), all three lines accumulated approximately 150-fold more squalene than WT (Figure 3b). As predicted, however, the abundance of dozens of other metabolites changed statistically significantly in all three engineered lines, and they encompassed multiple functional categories, including amino acids, carbohydrates, nucleotides and lipids (Figure 3c, Table S1).



Notably, when a metabolite's abundance was significantly changed in all three lines, the direction of change was generally identical (Figure 3c, Tables 1 and S1).

In both CN and C, approximately 120 metabolites changed significantly (Figure 3c, Table S1), and several accumulated more than 10-fold (Figure 3c, d, Table 1). Similarly, a transplastomic line in the Petit Havana (PH) background expressing only *SQS* significantly affected the abundance of nearly 100 metabolites (Table S2). Of these, the most notable was nicotine, which accumulated to nearly 11-fold higher levels than in untransformed plants (Table S2). Among those metabolites that accumulated in plants engineered in the WT 1068 background, three—tryptamine, tyramine and phenethylamine—increased without a corresponding increase in their aromatic amino acid precursors (Figure 3d, Table 1). Two TCA cycle-associated metabolites, citrate and succinate, accumulated to 15.2-fold and fourfold higher levels, respectively, in CN than in WT (Figure 3d, Table 1). In C, these metabolites accumulated to 5.4-fold and twofold higher levels than in WT (Figure 3d, Table 1). We also noted a concomitant reduction in photosynthetic Calvin cycle outputs fructose 6-phosphate (repressed up to fivefold) and glucose 6-phosphate (repressed up to 6.25-fold) (Table S1). In N, the abundance of approximately 65 metabolites changed statistically significantly (Figure 3c, Table S1), but with the exception of squalene (Figure 3b), they changed only two- to threefold at most (Figure 3c, Tables 1 and S1).

The impact of *FPS* and *SQS* expression on the transcription of nuclear genes

The metabolic changes and physiological phenotypes in all three lines prompted us to investigate transcriptional mechanisms that underlie these changes. Therefore, we performed poly(A)-selected RNA sequencing (RNA-seq) on leaf tissue from plants of each line. Leaf material was distinct from that used for the metabolomic analysis but was grown in an identical manner and harvested at an identical developmental stage, with transplastomic plants in the T0 generation. Importantly, poly(A) selection enriches for nuclear-encoded transcripts and depletes chloroplast transcripts, which lack an extended poly(A) tail. High-throughput sequencing reads were mapped to the tobacco genome, and transcripts were identified and quantified as previously described (Siervo *et al.*, 2014). Hierarchical clustering analysis of expression counts for predicted transcripts demonstrated that expression patterns of each replicate for the three different transgenic lines were similar (Figure 4a). PCA showed close clustering within lines (Figure 4b).

Unlike the metabolomic analysis, which showed N clustering more closely with WT and transplastomic CN and C lines clustering together (Figure 3a), transcriptomic analysis showed that overall gene expression profiles of each line were fairly separate. However, transcriptomes of engineered lines were more similar to each other than to WT (Figure 4b).

To annotate predicted transcripts, we first identified likely open reading frames (ORFs) and then used BLASTP with the NCBI *Nicotiana* database. This analysis led to many transcripts being assigned 'unknown' or 'uncharacterized' annotations or given no annotation. Additionally, multiple transcripts were often annotated to the same gene. Therefore, to increase the number of uniquely annotated ORFs, we used BLASTP with the TAIR10 version of the *Arabidopsis* transcriptome and assigned *Arabidopsis*-based annotations to the predicted transcripts. This analysis permitted more detailed functional categorization of each transcript and allowed the assignment of the most closely matched annotation to each transcript model.

At the 99.9% confidence level, more than twice as many transcripts were differentially expressed (DE) in CN: 19 076 compared with 7784 in N and 5224 in C. First, we found a 4300-fold increase in *FPS* expression and a 7200-fold increase in *SQS* expression in N compared with WT. In CN, *FPS* expression increased 3600-fold and nuclear transgenic *SQS* expression increased 1300-fold relative to WT. Next, we determined which transcripts were most highly changed in each line compared with WT, ranked them by overall fold change independent of the line(s), then identified the fold change in other two lines. Despite identifying DE transcripts in a line-independent manner, the majority of most highly induced (Table 2) and most highly repressed transcripts (Table 3) were common to all three lines and changed in the same direction. Strikingly, nearly half of most highly induced transcripts (9/20) were predicted to encode transporters of varying functions, including transporters for sucrose (SWEETs), phosphate, metals and auxin. In addition, transcripts with predicted roles in growth and development, metabolism and defence were induced. Consistent with these annotations, Gene Ontology (GO) analysis of biological process-related terms showed enrichment for terms related to stress and stimulus responses (Figure 4c). We also observed enrichment for transcripts involved in secondary metabolism (Figure 4c). GO analysis of cellular component- and molecular function-specific terms showed slight enrichment for photosynthesis-related components and processes (Figures S2 and S3).

Table 1 List of metabolites whose abundance changes at least 10-fold compared with wild-type (WT)

Metabolite	Category	Family	N/WT	CN/WT	C/WT
Squalene	Secondary metabolism	Terpenoids	154.10*	158.57*	158.86*
3-(2-piperidinyl)pyridine	Secondary metabolism	Alkaloids	-1.17	32.16*	31.82*
Pheophorbide A	Cofactors, prosthetic groups, electron carriers	Chlorophyll and haem metabolism	1.00	25.52*	2.58
Gluconate	Carbohydrate	Amino sugar and nucleotide sugar	1.55	24.85*	10.42*
Citrate	Carbohydrate	TCA cycle	1.13	14.16*	5.41*
Tryptamine	Amino acid	Aromatic amino acid metabolism (PEP derived)	1.76*	9.35*	13.48*
4-hydroxybutyrate (GHB)	Amino acid	Glutamate family (alpha-ketoglutarate derived)	-1.75	12.54*	-1.50
Tyramine	Amino acid	Aromatic amino acid metabolism (PEP derived)	-1.05	11.79*	11.21*
Phenethylamine	Amino acid	Aromatic amino acid metabolism (PEP derived)	1.57*	11.11*	11.55*
1-palmitoyl-GPI (16:0)	Lipid	Phospholipid	1.01	11.27*	3.93*

*Statistically significant difference in the accumulation of the indicated metabolite compared with WT.

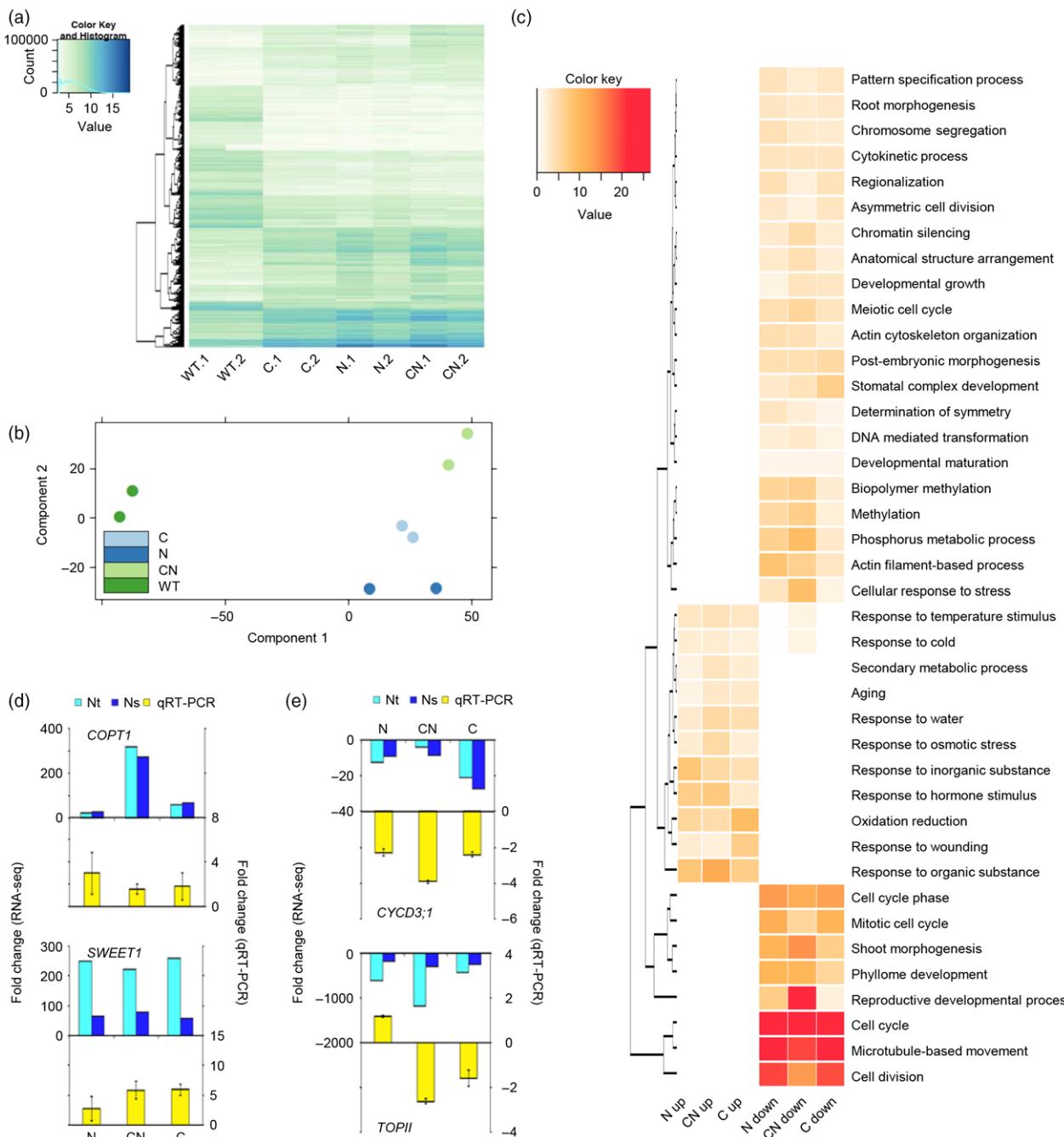


Figure 4 RNA-seq analysis of FPS/SQS plants. (a) Hierarchical clustering analysis and (b) PCA of duplicate samples. (c) Heat map showing overrepresented biological processes for up- and down-regulated transcripts in each line. Indicated values are $-\log_{10}$ of the Benjamini adjusted p-value. (d, e) RNA-seq data for transcripts derived from *Nicotiana tomentosiformis* (Nt, cyan) and *Nicotiana sylvestris* (Ns, blue) genomes (upper panels, primary y-axis) and qRT-PCR confirmation of transcript levels (lower panels, yellow, secondary y-axis) for transcripts that are induced (d) or repressed (e) relative to WT. In lower panels, error bars represent mean \pm SEM of two biological replicates with three technical replicates each. FPS, FARNESYL DIPHOSPHATE SYNTHASE; PCA, principal component analysis; SQS, SQUALENE SYNTHASE; WT, wild-type.

For a few transcripts, for example *SWEET1* (transcript 10), expression was induced similarly across all three lines (Table 2). However, for many other induced transcripts, although the largest change in expression differed, the greatest induction occurred most often in CN. For example, *COPT1* (transcript 6) and *PIN5* (transcript 8) were both induced approximately 300-fold in CN, whereas in N, *COPT1* was induced only 20-fold and *PIN5* was not significantly changed; and in C, *COPT1* was induced 70-fold

and *PIN5* was induced 50-fold. By contrast, only one transcript was most highly induced each in C (*Nicotiana tomentosiformis NCED4*, transcript 15) and N (*Nicotiana sylvestris TONOPLAST DICARBOXYLATE TRANSPORTER*, transcript 17), although the *N. tomentosiformis* 14-kDa protein (a lipid transporter, transcript 1) was also induced very highly in N—more than 800-fold (Table 2). Additionally, a subset of transcripts was induced to a similar level in both N and CN, including the *N. tomentosiformis*

Table 2 Top 20 overall most highly induced transcripts and their expression in other two transgenic/transplastomic lines, grouped by functional category. When possible, each locus includes an annotation for *Nicotiana tomentosiformis* (Nt) and an annotation for *Nicotiana sylvestris* (Ns). Values without asterisks indicate 99.9% confidence. Cells without values were not significantly changed for the indicated transcript in the indicated line

Rank	NCBI gene ID/At number	Nicotiana/Arabidopsis annotation	N/Wild-type (WT)	CN/WT	C/WT
Transporters					
1	LOC104087257/At4g12510	Nt 14 kDa PROLINE-RICH PROTEIN DC2.15-like/2S albumin superfamily protein	806.987	940.608	172.152
	LOC104242032/At4g12510	Ns 14 kDa PROLINE-RICH PROTEIN DC2.15-like/2S albumin superfamily protein	73.826	56.626	18.003
4	LOC104088673/At3g48740	Nt N3 SUGAR TRANSPORTER-like/SWEET11	87.235	201.150	65.779
	LOC104243017/At3g48740	Ns SWEET12-like/SWEET11	148.741	327.890	72.298
5	LOC104115159/At3g54700	Nt probable INORGANIC PHOSPHATE 1-7/PHT1;7	11.423*		15.762
	LOC104216800/At3g54700	Ns probable INORGANIC PHOSPHATE TRANSPORTER 1-2/PHT1;7	10.502*	323.984	299.250
6	LOC104086110/At5g59030	Nt COPPER TRANSPORTER 1-like/COPT1	20.421	318.867	55.805
	LOC104244096/At5g59030	Ns COPPER TRANSPORTER 1-like/COPT1	23.937	271.475	68.523
8	LOC104084825/At5g16530	Nt putative AUXIN EFFLUX CARRIER COMPONENT 8/PIN5		299.984	48.512
	LOC104249472/At5g16530	Ns putative AUXIN EFFLUX CARRIER COMPONENT 8/PIN5		201.077	32.503
10	LOC104086686/At1g21460	Nt SWEET1-like/SWEET1	248.285	220.603	259.497
	LOC104225662/At1g21460	Ns SWEET1-like/SWEET1	63.264	79.362	57.860
12	LOC104101946/At5g59520	Nt ZINC TRANSPORTER 2/ZIP2		32.226	
	LOC104211888/At5g59520	Ns ZINC TRANSPORTER 2/ZIP2	13.010	249.781	15.498
14	LOC104096904/At5g53190	Nt bidirectional sugar transporter SWEET3b/SWEET3		182.521	140.590
	LOC104241809/At5g53190	Ns bidirectional sugar transporter SWEET3/SWEET3		236.357	107.210
17	LOC104118881/At5g47560	Nt TONOPLAST DICARBOXYLATE TRANSPORTER/TDT	37.066	9.426	
	LOC104242854/At5g47560	Ns TONOPLAST DICARBOXYLATE TRANSPORTER/TDT	221.581	33.801	
Growth and development					
2	LOC104105402/At5g62040	Nt CEN-like 1/BFT		538.996	
	LOC104229471/At5g62040	Ns CEN-like 1/BFT		431.913	14.363
3	LOC104089336/At5g10625	Nt FLOWERING PROMOTING FACTOR 1-like 3/FPF1-like	115.961	356.184	16.807
	LOC104232534/At5g10625	Ns FLOWERING PROMOTING FACTOR 1-like 3/FPF1-like		440.531	12.179*
7	LOC104104199/At5g06720	Nt PEROXIDASE 15/PA2	20.544	312.576	80.377
	LOC104217681/At5g06720	Ns LIGNON-FORMING ANIONIC PEROXIDASE-like/PA2		81.758	
13	LOC104243032/At4g25820	Ns XYLOGLUCAN ENDOTRANSGLUCOSYLASE/HYDROLASE 24-like/XTR9	197.748	242.975	69.681
Secondary metabolism					
9	LOC104099208/At4g14090	Nt CROCETIN GLUCOSYLTRANSFERASE/ANTHOCYANIDIN-5-O-GLUCOSYLTRANSFERASE	94.367	272.367	18.904
	LOC104227958/At4g14090	Ns CROCETIN GLUCOSYLTRANSFERASE/ANTHOCYANIDIN-5-O-GLUCOSYLTRANSFERASE	9.567	12.620	4.611
15	LOC104104691/At4g19170	Nt CAROTENOID CLEAVAGE DIOXYGENASE 4/NCED4	131.189	101.045	234.293
19	LOC104107213/At4g15480	Nt CINNAMATE BETA-D-GLUCOSYLTRANSFERASE/UGT84A1	214.326	200.087	142.143
Hormone					
11	LOC104088764/At1g75750	Nt GIBBERELLIN-REGULATED 1-like/GASA1	58.475	256.662	11.149*
Primary metabolism					
16	LOC104096882/At5g48850	Nt uncharacterized/SDI1	38.021	227.743	
Cytoskeleton					
18	LOC104116945/At1g70810	Nt probable ADP-ribosylation factor GTPase-activating protein AGD11/CAR7	9.630	91.183	12.693
	LOC104249050/At1g70810	Ns probable ADP-ribosylation factor GTPase-activating protein AGD11/CAR7	52.144	216.566	32.270
Defence					
20	LOC104103886/At2g38870	Nt PROTEINASE INHIBITOR I-B/PR6-like	79.968	100.779	23.172
	LOC104231145/At2g38870	Ns PROTEINASE INHIBITOR I-A/PR6-like	131.969	210.761	35.884

*Value reported at the 99.0% confidence level.

14-kDa protein (transcript 1, 900-fold in CN and 800-fold in N) and both *N. tomentosiformis* and *N. sylvestris* PR6-like proteinase inhibitor (transcript 20, 100–200-fold in CN and ~100-fold in N) (Table 2).

Table 3 lists the 20 most highly repressed transcripts, with more than half (12/20) predicted to function in the cell cycle, cell wall remodelling and DNA remodelling, and additional functional

categories included kinases, metabolism and growth and development. GO analysis for 'biological process'-specific terms confirmed substantial enrichment among repressed transcripts for those with predicted functions in the cell cycle (Figure 4c). GO analysis for 'cellular component'-related terms showed enrichment for cell wall, chromosomes and cytoskeletal components among down-regulated transcripts (Figure S2), and GO analysis

Table 3 Top 20 overall most highly repressed transcripts and their expression in other two transgenic/transplastomic lines, grouped by functional category. When possible, each locus includes an annotation for *Nicotiana tomentosiformis* (Nt) and an annotation for *Nicotiana sylvestris* (Ns)

Rank	NCBI gene ID/At number	Nicotiana annotation/Arabidopsis annotation	N/Wild-type (WT)	CN/WT	C/WT
Cell cycle					
1	LOC10409419/At1g30690	Nt <i>PATELLIN-4</i> /Sec14-like	−816.166	−263.187	−2262.789
	LOC10421949/At1g30690	Ns <i>PATELLIN-4</i> /Sec14-like	−493.844	−746.213	−666.637
7	LOC10408601/At5g06150	Nt <i>CYCLIN S13-7</i> -like/CYCB1;2	−726.831	−364.512	−559.138
	LOC10423757/At5g06150	Ns <i>CYCLIN S13-7</i> -like/CYCB1;2	−1052.573	−524.124	−1325.489
10	LOC10421173/At4g11080	Ns <i>HIGH MOBILITY GROUP B 6</i> -like/3XHMG-BOX1	−531.236	−317.492	−1124.502
11	LOC10424297/At1g44110	Ns <i>CYCLIN-A1-4</i> /CYCA1;1	−257.417	−401.190	−989.323
12	LOC10409411/At3g20150	Nt kinesin-like <i>KIN12A</i> /kinesin family protein	−732.380	−361.282	−918.263
	LOC10424399/At3g20150	Ns kinesin-like <i>KIN12A</i> /kinesin family protein	−368.772	−322.527	−453.958
13	LOC10411222/At1g03780	Nt <i>TPX2</i> -like/TPX2	−584.023	−504.647	−308.380
	LOC10421712/At1g03780	Ns <i>TPX2</i> -like/TPX2	−915.733	−790.737	−694.173
15	LOC10410080/At2g26760	Nt <i>CYCLIN S13-7</i> -like/CYCB1;4	−682.727	−593.531	−850.024
	LOC10422394/At2g26760	Ns <i>CYCLIN S13-7</i> -like/CYCB1;4	−248.516	−119.077	−306.997
16	LOC10409883/At1g08560	Nt <i>KNOLLE</i> /KNOLLE	−654.447	−125.074	−298.701
	LOC10422292/At1g08560	Ns <i>KNOLLE</i> /KNOLLE	−647.369	−163.412	−847.873
Cell wall/cuticle					
3	LOC10411994/At4g22010	Nt <i>L-ASCORBATE OXIDASE</i> homologue/SKS4	−459.175	−660.733	−594.408
	LOC10423150/At4g22010	Ns <i>L-ASCORBATE OXIDASE</i> homologue/SKS4	−481.000	−1001.534	−1473.924
6	LOC10410009/At5g47500	Nt probable <i>PECTINESTERASE 68</i> /PME5	−642.662	−370.989	−1355.422
	LOC10421908/At5g47500	Ns probable <i>PECTINESTERASE 68</i> /PME5	−345.366	−518.392	−831.930
19	LOC10409611/At1g02205	Nt <i>ECERIFERUM 1</i> -like/CER1	−38.939	−91.297	−78.052
	LOC10424030/At1g02205	Ns <i>ECERIFERUM 1</i> -like/CER1	−199.464	−546.898	−799.481
Kinase					
4	LOC10411606/At3g51740	Nt probably inactive leucine-rich repeat receptor-like protein kinase IMK2/IMK2	−1446.626	−764.462	−457.671
	LOC10423705/At3g51740	Ns probable leucine-rich repeat receptor-like protein kinase IMK3/IMK3	−305.037	−265.833	−219.477
17	LOC10409992/At5g43020	Nt probable inactive receptor kinase At5g67200/LRR kinase family	−168.644	−572.847	−841.587
	LOC10423543/At5g43020	Ns probable inactive receptor kinase At5g67200/LRR kinase family	−32.928	−96.819	−44.635
Miscellaneous					
18	LOC10409693/At4g31840	Nt <i>EARLY NODULIN-LIKE 1</i> /ENODL15	−395.968	−118.288	−821.479
	LOC10421187/At4g31840	Ns <i>EARLY NODULIN-LIKE 3</i> /ENODL15	−343.122	−120.464	−623.076
20	LOC10412083/At5g23420	Nt <i>HIGH MOBILITY GROUP B 7</i> -like/HMGB6	−500.882	−134.517	−476.056
	LOC10424371/At5g23420	Ns <i>HIGH MOBILITY GROUP B 7</i> -like/HMGB6	−262.125	−222.445	−790.240
Lipid metabolism					
2	LOC10412153/At5g33370	Nt GDSL esterase/lipase At5g33370-like/GDSL-motif esterase/lipase	−546.326	−2106.034	−1221.907
	LOC10422041/At5g33370	Ns GDSL esterase/lipase At5g33370-like/GDSL-motif esterase/lipase	−79.807	−93.596	−224.898
Secondary metabolism					
5	LOC10408627/At5g66230	Nt uncharacterized/chalcone-flavanone isomerase	−828.169	−136.904	−490.657
	LOC10421950/At5g66230	Ns uncharacterized/chalcone-flavanone isomerase	−686.380	−310.377	−1419.234
Growth and development					
8	LOC10409319/At2g42840	Nt <i>PROTODERMAL FACTOR 1</i> -like/PDF1	−1003.212	−869.240	−1255.300
	LOC10422232/At2g42840	Ns <i>PROTODERMAL FACTOR 1</i> -like/PDF1	−486.103	−145.233	−302.012
DNA remodelling					
9	LOC10412078/At3g23890	Nt <i>DNA TOPOISOMERASE II</i> -like/TOPII	−615.688	−1172.774	−442.749
	LOC10424672/At3g23890	Ns <i>DNA TOPOISOMERASE II</i> -like/TOPII	−189.431	−287.318	−247.807
Cytoskeleton					
14	LOC10408502/At1g03470	Nt uncharacterized/KIP1-like	−688.519	−235.486	−866.163
	LOC10422085/At1g03470	Ns uncharacterized/KIP1-like	−306.668	−252.734	−287.424

for 'molecular function'-related terms showed enrichment for transcripts predicted to be involved in binding nucleosides and cytoskeletal proteins (Figure S3).

Without exception, each of the most highly repressed transcripts were repressed across all three lines, and there were some—for example the *N. tomentosiformis* TPX2-like gene (transcript 13) and the *N. sylvestris* KIP1-family gene (transcript 14)—whose expression was comparable in all three lines. However, there were other

transcripts for which the magnitude of change did vary. Strongest repression was often found in C, as in the case of the *N. tomentosiformis* *PATELLIN-4* (transcript 1, repressed more than 2000-fold) and the *N. sylvestris* *HMG PROTEIN6* (transcript 10, repressed 1000-fold). In other instances, as for a transcript annotated to a *N. tomentosiformis* GDSL-motif esterase/lipase (transcript 2) and for a *N. tomentosiformis* *DNA TOPOISOMERASE II* (transcript 9), CN clearly showed the strongest repression: 2000-fold and

1100-fold, respectively. CN also showed a unique enrichment for down-regulated transcripts predicted to be involved in reproductive development (Figure 4c). The N line also had a large effect on the expression of certain transcripts, most notably that for the *N. tomentosiformis* IMK2 kinase (transcript 4, repressed 1400-fold). Two other transcripts, *N. sylvestris* *CYCLIN B1;2* and *N. tomentosiformis* *PDF1* (transcripts 7 and 8, respectively) were also repressed more than 1000-fold in N (Table 3). Interestingly, in cases in which a transcript was most highly repressed in two lines, the two with the strongest effect tended to be N and C (e.g. for *N. sylvestris* *CYCLIN B1;2* and *N. tomentosiformis* kinesin *KIN12A*).

To further confirm the validity of RNA-seq data, we performed qRT-PCR for select induced and repressed transcripts using biologically independent samples at an identical developmental stage. All transcripts predicted to be induced by RNA-seq were verified to be induced by qRT-PCR (Figure 4d), and with one exception (*TOPII* in N), transcripts predicted by RNA-seq to be repressed were also repressed relative to WT in qRT-PCR (Figure 4e). Plotting the \log_2 fold change of each transcript using qRT-PCR versus RNA-seq showed a good correlation between results from both methods, with R values ranging from 0.690 to 0.899 (Figure S4).

Discussion

To date, few studies on plant metabolic engineering have looked beyond the pathway of interest to assess unanticipated consequences, and no report has compared metabolomic and transcriptomic effects of expressing transgenes from different organellar genomes. Here, we demonstrate that expressing chloroplast-targeted *FPS* and/or *SQS* from the nuclear genome (N line), the chloroplast genome (C line) or both (CN line) has broad off-target effects on the metabolome and transcriptome. However, because many of the same metabolites and transcripts are changed in a similar direction across all three lines, *FPS* and *SQS* expression may cause broad but somewhat predictable off-target effects.

Because the chloroplast genome promotes higher transgene expression than the nuclear genome (Verma and Daniell, 2007), one goal of this study was to increase squalene production by transplastomically expressing codon-optimized *FPS* and/or *SQS*. However, neither transplastomic line accumulated significantly more squalene than N (Figure 3b, Tables 1 and S1), suggesting a substrate or storage limit. Consistent with this, in the background of PH, which has fewer glandular trichomes than 1068 (Wu et al., 2012), we could generate plants expressing *SQS* only; transforming PH chloroplasts with pLD-FPS-SQS yielded shoots that did not survive, and similar problems were encountered using nuclear transformation constructs that targeted *FPS* and *SQS* to chloroplasts. Inadequate trichomes (sinks) in PH may have caused this, perhaps through feedback inhibition. Indeed, the 2-C-methyl-D-erythritol 4-phosphate (MEP) pathway of terpene biosynthesis is under feedback inhibition by its own products, dimethylallyl diphosphate (DMAPP) and isopentenyl diphosphate (IPP) (Banerjee et al., 2013), and IPP is the substrate of *FPS*. One strategy to increase squalene could be to increase sink capacity by sequestering squalene in storage compartments such as lipid droplets (Wang et al., 2015) or additional trichomes (Lange, 2015). Trichome density could be increased via transgenic expression of β -glucosidase (Jin et al., 2011), which improves artemisinin yields in transgenic *Artemisia annua* (Singh et al., 2015).

Among the 65–120 metabolites changed across engineered lines, in transplastomic lines, we noted an increase in TCA cycle-related metabolites, including citrate and succinate (Figure 3d), and a reduction in fructose 6-phosphate and glucose 6-phosphate (Table S1). Fructose 6-phosphate and glucose 6-phosphate are both photosynthetic outputs from the Calvin cycle and are synthesized from glyceraldehyde 3-phosphate, a three-carbon Calvin cycle output. In chloroplasts, squalene biosynthesis is initiated by the conversion of glyceraldehyde 3-phosphate and pyruvate into 1-deoxy-D-xylulose-5-phosphate in a committed step catalysed by 1-deoxy-D-xylulose-5-phosphate synthase. Therefore, the 150-fold increase in squalene levels and the decrease in fructose 6-phosphate and glucose 6-phosphate biosynthesis observed in engineered lines suggest that glyceraldehyde 3-phosphate is repartitioned from the synthesis of sugars into the production of terpenes (Wang et al., 2015).

The chloroplast amino acid pool is limited (Bally et al., 2009), and chloroplast transgene products are produced at the expense of resident proteins such as Rubisco (Bally et al., 2011; Ruhlman et al., 2010). Therefore, given the accumulation of Flag-SQS in transplastomic lines (Figure 1d), it is not surprising that amino acid metabolism was altered in CN and C (Figure 3c and d, Table 1). Decarboxylated aromatic amino acids accumulated to at least eightfold higher levels in CN and C than in WT without an increase in their corresponding precursors (Figure 3d, Table 1). This result demonstrates that transplastomic lines are not impaired in amino acid biosynthesis and suggests the diversion of carbon into squalene biosynthesis, leading to an imbalance between carbon and nitrogen that explains the accumulation of these amino acid derivatives. Additionally, the abundance of various nucleotides was affected (Figure 3c; Table S1), possibly due to large quantities of *FPS* and/or *SQS* mRNA.

Transcriptomic analysis revealed thousands of DE transcripts between each line and WT (Figure 4a). Induction of transporter-related transcripts (Table 2) may reflect the need to move additional metabolites from sites of synthesis to sites of use or storage. Additionally, highly repressed cell cycle transcripts (Table 3, Figure 4c) correlated well with growth phenotypes (Figure 2c). Importantly, all most highly DE transcripts were changed in the same direction across all three lines (Tables 2 and 3, Figure 4d,e), regardless of the compartment(s) from which transgenes were expressed.

We also observed differences in transcriptomes that varied with the compartment(s) from which transgenes were expressed, including the number and magnitude of change of DE transcripts. Importantly, transcriptomic data also showed no correlation between transgene expression and the effect on the expression of nuclear-encoded genes; the N line expressed up to 4000-fold less Flag-SQS than transplastomic lines (Figure 1d) but affected more nuclear-encoded transcripts than C, and it had the strongest influence on the expression of certain transcripts (Tables 2 and 3). Additionally, when a transcript was most strongly affected in two lines, one of those lines was generally N. Therefore, alterations in gene expression are likely the result of metabolites and not transgene products. Indeed, levels of Flag-SQS did not vary much between transplastomic lines, but the effect on nuclear-encoded genes was quite significant.

The number of DE transcripts reported here is considerably greater than those reported in previous studies (Ricroch et al., 2011). For example, transgenic rice expressing choline oxidase showed only 165 DE transcripts between engineered and parental lines (Kathuria et al., 2009). In transgenic potatoes with

elevated or knocked down expression of *sucrose synthase*, 50 and 357 DE genes were found, respectively (Baroja-Fernández *et al.*, 2009). In transgenic rice expressing *anthranilate synthase*, after correcting for variation between samples, only 22 genes met authors' criteria for differential expression (Dubouzet *et al.*, 2007). Compared with DNA microarrays, the greater sensitivity of RNA-seq may facilitate the detection of far more transcriptomic changes (Wang *et al.*, 2009), and transgene expression from different cellular compartments likely also increased the number of DE transcripts (discussed below).

One potential explanation for strong phenotypes observed in CN is disparate levels of *FPS* and *SQS* transgene products. Transplastomic CN and C lines accumulate at least 2000-fold more Flag-SQS than N (Figure 1d), and expression of *FPS* from the nuclear genome alone may not produce enough transgene product to match levels of *SQS* in CN (Figure 1c). Therefore, phenotypes observed in CN may be mitigated by incorporating *FPS* into the chloroplast genome. Similarly, in transplastomic PH plants expressing only *SQS*, the abundance of dozens of metabolites changed by >10-fold (Table S2). Notably, one of these metabolites was nicotine, which accumulated to nearly 11-fold higher levels than in untransformed plants (Table S2), supporting the idea that metabolic imbalance can result in substantial off-target effects on entirely unrelated pathways.

Because of transgene compartmentalization, transplastomic plants are excellent tools to study retrograde signalling. In CN, because the parental N line only contained approximately 8000 DE transcripts, additional 11 000 DE transcripts may have changed due to retrograde signalling. In the C line, because transgenes and resulting products are totally contained within chloroplasts, all changes in the expression of the more than 5000 nuclear-encoded transcripts must be the result of retrograde signalling.

Although our data do not permit us to identify the retrograde signal, it is likely not squalene; otherwise, similar squalene levels (Figure 3b) would cause similar changes in gene expression. However, not only the number of DE transcripts differed, but also their expression levels (Tables 2 and 3). The signal may therefore be a different metabolite or metabolites, perhaps one affected by the imbalance between *SQS* and *FPS* levels in CN (Figure 3c, Table S1), but investigation of this possibility is beyond the scope of this study. Interestingly, an upstream molecule in the MEP pathway, methylerythritol cyclodiphosphate (MEcPP), acts as a retrograde signalling molecule under high light and wounding; these stresses induce accumulation of MEcPP to ~two- to threefold higher levels than in controls (Xiao *et al.*, 2012). Altered flux through the MEP pathway may change the abundance of MEcPP, thereby causing at least a subset of observed transcriptomic changes in engineered plants. The effect of *FPS* and *SQS* expression could extend downstream of MEcPP and parallel to farnesyl diphosphate (FPP) and/or squalene, and these metabolites may also serve as retrograde signals. Both IPP and DMAPP are precursors for the synthesis of carotenoids such as β -carotene (Nisar *et al.*, 2015), and a volatile β -carotene derivative, β -cyclocitral, has been proposed as a stress-induced retrograde signalling molecule (Ramel *et al.*, 2012). Increased demand for IPP by *FPS* or for FPP by *SQS* may affect carotenoid metabolism, and, consequently, β -cyclocitral levels, causing altered retrograde signalling. Although none of these metabolites were significantly changed between WT and engineered lines in our analysis, it does not exclude the possibility that the abundance of one or more of them was altered, leading to

changes in phenotype and gene expression. If such a difference had occurred, small changes needed to induce signalling (e.g. for MEcPP) and differences in detection and analytical methods may have excluded these metabolites from analysis, especially if differences were statistically insignificant.

Gene products have also been suggested to act as highly specific retrograde signalling molecules, and transgenes engineered via the chloroplast genome that are unrelated to metabolism can influence the expression of nuclear genes. For example, chloroplast expression of *Arabidopsis TIC40* or γ -tocopherol methyltransferase has been shown to promote massive proliferation of the chloroplast inner membrane and the up-regulation of associated nuclear-encoded inner membrane proteins without affecting nuclear-encoded proteins that are targeted to the outer or thylakoid membranes (Jin and Daniell, 2014; Singh *et al.*, 2008). Additionally, proteins have been shown to be released from intact chloroplasts (Kwon *et al.*, 2013b), providing yet another observation in support of the idea that proteins themselves can be retrograde signals. Therefore, the retrograde signal(s) acting in this case could be a metabolite or a gene product.

Previous studies on genetically modified plants have reported that environmental factors and cultivar-specific differences play larger roles in altering the transcriptome and metabolome than does transgene expression (Baker *et al.*, 2006; Clarke *et al.*, 2013; Kogel *et al.*, 2010; Ricroch *et al.*, 2011). *FPS* and *SQS* expression may cause clear transcriptomic, metabolomic and phenotypic changes because they are metabolism related. For example, transplastomic tobacco expressing twelve genes for the biosynthesis of artemisinic acid (AA) accumulated only approximately 0.1 mg/g fresh weight (FW) AA, but growth was still reduced (Saxena *et al.*, 2014), and nuclear transgenic plants accumulating linalool and nerolidol accumulated 1.5 μ g/g FW of these terpenes but still displayed delayed growth (Aharoni *et al.*, 2003). These observations may reflect the partitioning of carbon away from biomass and into the desired metabolite (Melis, 2013). By contrast, transgenic/transplastomic expression of genes that are unrelated to metabolism has few to no phenotypic effects and, in the case of transplastomic plants, can often be rescued by adding exogenous nitrogen (Bally *et al.*, 2009, 2011; De Cosa *et al.*, 2001; Ruhlman *et al.*, 2010; Viitanen *et al.*, 2004). Therefore, metabolite engineering and not protein engineering likely accounts for these phenotypes.

Using cutting-edge -omics technology, we show that metabolic engineering via the nuclear and/or chloroplast genomes can result in broad off-target effects in the metabolome and transcriptome. However, these effects may be predictable and can therefore be minimized or exploited. By focusing on global consequences of metabolic engineering rather than simply searching for another individual signalling molecule, our results provide a unique, holistic view of metabolite-mediated intercompartmental signalling that can be used as a framework for future studies on both metabolic engineering and metabolite-mediated anterograde and retrograde signalling.

Experimental procedures

Vector construction, plant transformation and characterization via PCR, Southern blot, northern blot and western blot

The WT 1068 introduction and the N nuclear transgenic line were as previously described (Wu *et al.*, 2012). All N plants were

homozygous siblings derived from the same transformation event. To construct the vectors used to generate transplastomic lines, DNA sequences encoding Flag-tagged yeast SQS (GenBank accession NM001179321) and 4xHis-tagged avian FPS (GenBank accession P08836) were codon-optimized for enhanced chloroplast expression (Daniell *et al.*, 2009) and synthesized by Integrated DNA Technologies (Coralville, IA). The optimized *Flag-SQS* coding sequence was cloned into the *Nde*I and *Xba*I sites of pLD-ctv under the control of the *psbA* promoter, 5'-UTR and 3'-UTR (Daniell *et al.*, 2005). To construct pLD-FPS-SQS, an intermediate vector was assembled that contained the *Prnr-g10/His-FPS/Trbcl* cassette. This vector was digested with *Sal*I, and the released cassette was ligated into *Sal*I-digested pLD-SQS. Both vectors were confirmed by sequencing. Chloroplast transformation of WT with pLD-FPS-SQS and of N with pLD-SQS was performed as described previously (Verma *et al.*, 2008). PCR and western blot of putative transplastomic plants were carried out as described (Kwon *et al.*, 2013a; Verma *et al.*, 2008). Western blot was performed using an anti-DYKDDDDK antibody (Life-Tein, South Plainfield, NJ) at a 1 : 1000 dilution and a horseradish peroxidase-conjugated anti-mouse secondary antibody (Southern Biotech, Birmingham, AL, USA) at a 1 : 4000 dilution. Densitometric analysis was carried out using the gel analysis feature of ImageJ, Bethesda, MD, USA. Southern blot was performed using the DIG High Prime DNA Labeling and Detection Kit (Roche Diagnostics, Indianapolis, IN) according to the manufacturer's instructions. In brief, 0.5 µg of total tobacco DNA was digested with *Af*III or *Hind*III and resolved on a 0.8% agarose gel. The DNA was blotted onto a positively charged nylon membrane (Nytran SPC; GE Healthcare, Marlborough, MA). Hybridization with DIG-labelled probe for the *trnl-trnA* flanking region (Figure 1a) was conducted at 41 °C in a UVP HB-1000 hybridizer (UVP LLC, Upland, CA), and signals were detected with CSPD substrate and X-ray film. Probe for northern blot analysis was synthesized using the DIG High Prime DNA Labeling and Detection Kit with a PCR product containing the *rbcL* 3'-UTR. Northern blot was performed using total RNA (see below) run on a 0.9% denaturing agarose gel. The RNA was blotted onto a Nytran membrane and hybridized with the probe at 42 °C. Signals were detected as for the Southern blot. For phenotypic analysis of homoplasmic transplastomic lines, data are representative of at least two independent transformation events.

Metabolome analysis

Samples were harvested from the third leaf of plants of the same developmental stage grown *via* nodal cutting in sterile tissue culture conditions. Harvested tissue was lyophilized in a Genesis lyophilizer (SP Scientific, Warminster, PA), and quadruplicate samples were sent to Metabolon (Durham, NC). Samples were analysed using a platform consisting of GC/MS and two UHPLC/MS/MS analyses, one optimized for acidic species and one for basic species (Clarke *et al.*, 2013; Evans *et al.*, 2009). Metabolomic data are available in Tables S1 and S2.

RNA sequencing and analysis and GO analysis

mRNA-seq was performed as previously described (Elliott *et al.*, 2013). Briefly, total RNA was purified from the third youngest leaf of two plants from each line at the same developmental stage grown in sterile tissue culture conditions using a miRNeasy kit (Qiagen, Valencia, CA). Poly(A)+ RNA was isolated using oligo(dT) beads (Thermo Fisher Scientific, Waltham, MA). RNA was

fragmented for 7 min using Fragmentation Reagent (Thermo Fisher Scientific). mRNA-seq libraries were generated using an Illumina mRNA-seq kit (Illumina, San Diego, CA). Sequencing was performed at the University of Pennsylvania Next Generation Sequencing Core. Reads were trimmed with Cutadapt and mapped to the tobacco genome with Tophat2 (Kim *et al.*, 2013; Martin, 2011; Sierro *et al.*, 2014; Trapnell *et al.*, 2009). Cufflinks was used to predict transcripts (Trapnell *et al.*, 2010), HTSeq was used to quantify expression of each transcript (Anders *et al.*, 2015), and DESeq2 was used to perform differential expression analysis and data normalization (Love *et al.*, 2014). We used TransDecoder (Haas *et al.*, 2013) to identify ORFs within predicted transcripts and annotated them to TAIR10 proteins. RNA-seq data have been deposited in GEO under accession number GSE74103.

To perform GO analysis, transcripts were separated into up- and down-regulated and then merged with identical TAIR10 annotations. These gene lists were used as input for the DAVID functional annotation tool (Huang *et al.*, 2009). Heat maps were generated using only level 2 and 3 terms with Benjamini adjusted p-values < 0.05.

Total RNA extraction, cDNA synthesis and qRT-PCR

Total RNA was isolated from the third youngest leaf of two plants per line grown in sterile tissue culture conditions using a PureLink RNA Mini Kit (Thermo Fisher Scientific) with on-column DNase treatment. Three micrograms of total RNA was reverse-transcribed using random hexamers and a Reverse Transcription System (Promega, Madison, WI) according to the manufacturer's recommendations. The cDNA was diluted 50-fold in nuclease-free water, and 6.5 µL of diluted cDNA was used as template for qRT-PCRs.

qRT-PCR was performed in a volume of 20 µL using Power SYBR Green PCR master mix (Thermo Fisher Scientific). Primers used for RT-PCR analysis are listed in Table S3. Reactions were run on a StepOnePlus Real-Time PCR system (Thermo Fisher Scientific). Calculations were performed using the $2^{\Delta\Delta Ct}$ method and normalized to *ACT* and *EF1a*.

Acknowledgements

Authors acknowledge contributions from Drs. Dheeraj Verma, Anirudha Dixit, Man Zhou and John Cupp in the Daniell laboratory during early stages of this project. We thank Dr. Joe Chappell (University of Kentucky) for N seeds and Drs. Danny Alexander and Lining Guo (Metabolon, Inc.) for contributions to the metabolomic study. This work was supported in part by a grant from the U.S. Department of Energy ARPA-E programme to H.D. and J.S.Y. and NIH grants R01 HL107904 and R01 HL109442 to H.D.

Author contributions

E.K.P. performed the research and analysed/interpreted data in Figures 1c and 2–4, Tables 1–3 and the supplementary data and wrote this manuscript. J.S. contributed data in Figure 1b,d and related text in Results and Methods sections. I.M.S., S.J.G. and B.D.G. analysed/interpreted data in Figure 4, Tables 2 and 3 and Figures S2–S4 and contributed related text in Results and Materials and methods sections. J.S.Y. analysed/interpreted data and contributed text in the Introduction and Discussion sections. H.D.

conceived and designed the study, analysed/interpreted data and wrote this manuscript.

Conflict of interest

H.D. is an inventor in several U.S. and international patents related to chloroplast transformation technology.

References

Aharoni, A., Giri, A.P., Deuerlein, S., Griepink, F., de Kogel, W.J., Verstappen, F.W., Verhoeven, H.A. et al. (2003) Terpenoid metabolism in wild-type and transgenic *Arabidopsis* plants. *Plant Cell*, **15**, 2866–2884.

Anders, S., Pyl, P.T. and Huber, W. (2015) HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.

Baker, J.M., Hawkins, N.D., Ward, J.L., Lovegrove, A., Napier, J.A., Shewry, P.R. and Beale, M.H. (2006) A metabolomic study of substantial equivalence of field-grown and genetically modified wheat. *Plant Biotechnol. J.*, **4**, 381–392.

Bally, J., Nadai, M., Vitel, M., Rolland, A., Dumain, R. and Dubald, M. (2009) Plant physiological adaptations to the massive foreign protein synthesis occurring in recombinant chloroplasts. *Plant Physiol.*, **150**, 1474–1481.

Bally, J., Job, C., Belghazi, M. and Job, D. (2011) Metabolic adaptation in transplastomic plants massively accumulating recombinant proteins. *PLoS ONE*, **6**, e25289.

Banerjee, A., Wu, Y., Banerjee, R., Li, Y., Yan, H. and Sharkey, T.D. (2013) Feedback inhibition of deoxy-D-sugulose-5-phosphate synthase regulates the methylerythritol 4-phosphate pathway. *J. Biol. Chem.*, **288**, 16926–16936.

Baroja-Fernández, E., Muñoz, F.J., Montero, M., Etxeberria, E., Sesma, M.T., Ovecka, M., Bajahí, A. et al. (2009) Enhancing sucrose synthase activity in transgenic potato (*Solanum tuberosum* L.) tubers results in increased levels of starch, ADPGlucose and UDPGlucose and total yield. *Plant Cell Physiol.*, **50**, 1651–1662.

Bobik, K. and Burch-Smith, T.M. (2015) Chloroplast signaling within, between and beyond cells. *Front. Plant Sci.*, **6**, 781.

Chan, H.-T. and Daniell, H. (2015) Plant-made oral vaccines against human infectious diseases—are we there yet? *Plant Biotechnol. J.*, **13**, 1056–1070.

Chi, W., Feng, P., Ma, J. and Zhang, L. (2015) Metabolites and chloroplast retrograde signaling. *Curr. Opin. Plant Biol.*, **25**, 32–38.

Clarke, J.D., Alexander, D.C., Ward, D.P., Ryals, J.A., Mitchell, M.W., Wulff, J.E. and Guo, L. (2013) Assessment of genetically modified soybean in relation to natural variation in the soybean seed metabolome. *Sci. Rep.*, **3**, 3082.

Daniell, H., Ruiz, O.N. and Dhingra, A. (2005) Chloroplast genetic engineering to improve agronomic traits. *Methods Mol. Biol.*, **286**, 111–138.

Daniell, H., Ruiz, G., Denes, B., Sandberg, L. and Langridge, W. (2009) Optimization of codon composition and regulatory elements for expression of human insulin like growth factor-1 in transgenic chloroplasts and evaluation of structural identity and function. *BMC Biotechnol.*, **9**, 33.

De Cosa, B., Moar, W., Lee, S.B., Miller, M. and Daniell, H. (2001) Overexpression of the Bt cry2Aa2 operon in chloroplasts leads to formation of insecticidal crystals. *Nat. Biotechnol.*, **19**, 71–74.

Dhingra, A., Portis, A.R. Jr and Daniell, H. (2004) Enhanced translation of a chloroplast-expressed RbcS gene restores small subunit levels and photosynthesis in RbcS antisense plants. *Proc. Natl. Acad. Sci. USA*, **101**, 6315–6320.

Dubouzet, J.G., Ishihara, A., Matsuda, F., Miyagawa, H., Iwata, H. and Wakasa, K. (2007) Integrated metabolomic and transcriptomic analyses of high-tryptophan rice expressing a mutant anthranilate synthase alpha subunit. *J. Exp. Bot.*, **58**, 3309–3321.

Elliott, R., Li, F., Dragomir, I., Chua, M.M., Gregory, B.D. and Weiss, S.R. (2013) Analysis of the host transcriptome from demyelinating spinal cord of murine coronavirus-infected mice. *PLoS ONE*, **8**, e75346.

Estavillo, G.M., Crisp, P.A., Pornsiriwong, W., Wirtz, M., Collinge, D., Carrie, C., Giraud, E. et al. (2011) Evidence for a SAL1-PAP chloroplast retrograde pathway that functions in drought and high light signaling in *Arabidopsis*. *Plant Cell*, **23**, 3992–4012.

Evans, A.M., DeHaven, C.D., Barrett, T., Mitchell, M. and Milgram, E. (2009) Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems. *Anal. Chem.*, **81**, 6656–6667.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B. et al. (2013) De novo sequence reconstruction from RNA-Seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.*, **8**, 1494–1512.

Huang, D.W., Sherman, B.T. and Lempicki, R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.

Jensen, P.E. and Leister, D. (2014) Chloroplast evolution, structure and functions. *F1000 Prime Rep.*, **6**, 40.

Jin, S. and Daniell, H. (2014) Expression of γ -tocopherol methyltransferase in chloroplasts results in massive proliferation of the inner envelope membrane and decreases susceptibility to salt and metal-induced oxidative stresses by reducing reactive oxygen species. *Plant Biotechnol. J.*, **12**, 1274–1285.

Jin, S. and Daniell, H. (2015) The engineered chloroplast genome just got smarter. *Trends Plant Sci.*, **20**, 622–640.

Jin, S., Kanagaraj, A., Verma, D., Lange, T. and Daniell, H. (2011) Release of hormones from conjugates: chloroplast expression of β -glucosidase results in elevated phytohormone levels associated with significant increase in biomass and protection from aphids or whiteflies conferred by sucrose esters. *Plant Physiol.*, **155**, 222–235.

Kathuria, H., Giri, J., Nataraja, K., Murata, N., Udayakumar, M. and Tyagi, A. (2009) Glycinebetaine-induced water-stress tolerance in codA-expressing transgenic indica rice is associated with up-regulation of several stress responsive genes. *Plant Biotechnol. J.*, **7**, 512–526.

Kim, S.-K. and Karadeniz, F. (2012) Biological importance and applications of squalene and squalane. *Adv. Food Nutr. Res.*, **65**, 223–233.

Kim, D., Perte, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**, R36.

Kogel, K.-H., Voll, L.M., Schäfer, P., Jansen, C., Wu, Y., Langen, G., Imani, J. et al. (2010) Transcriptome and metabolome profiling of field-grown transgenic barley lack induced differences but show cultivar-specific variances. *Proc. Natl. Acad. Sci. USA*, **107**, 6198–6203.

Kumar, S., Hahn, F.M., Baidoo, E., Kahlon, T.S., Wood, D.F., McMahan, C.M., Cornish, K. et al. (2012) Remodeling the isoprenoid pathway in tobacco by expressing the cytoplasmic mevalonate pathway in chloroplasts. *Metab. Eng.*, **14**, 19–28.

Kwon, K.C., Nityanandam, R., New, J.S. and Daniell, H. (2013a) Oral delivery of bioencapsulated exendin-4 expressed in chloroplasts lowers blood glucose level in mice and stimulates insulin secretion in beta-TG6 cells. *Plant Biotechnol. J.*, **11**, 77–86.

Kwon, K.C., Verma, D., Jin, S., Singh, N.D. and Daniell, H. (2013b) Release of proteins from intact chloroplasts induced by reactive oxygen species during biotic and abiotic stress. *PLoS ONE*, **8**, e67106.

Lange, B.M. (2015) The evolution of plant secretory structures and emergence of terpenoid chemical diversity. *Annu. Rev. Plant Biol.*, **66**, 139–159.

Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10–12.

Melis, A. (2013) Carbon partitioning in photosynthesis. *Curr. Opin. Chem. Biol.*, **17**, 453–456.

Nisar, N., Li, L., Lu, S., Khin, N.C. and Pogson, B.J. (2015) Carotenoid metabolism in plants. *Mol. Plant*, **8**, 68–82.

Nott, A., Jung, H.S., Koussevitzky, S. and Chory, J. (2006) Plastid-to-nucleus retrograde signaling. *Annu. Rev. Plant Biol.*, **57**, 739–759.

O'Hagan, D.T., Rappuoli, R., De Gregorio, E., Tsai, T. and Del Giudice, G. (2011) MF59 adjuvant: the best insurance against influenza strain diversity. *Expert Rev. Vaccines*, **10**, 447–462.

Ramel, F., Birtic, S., Ginies, C., Soubigou-Taconnat, L., Triantaphylidès, C. and Havaux, M. (2012) Carotenoid oxidation products are stress signals that mediate gene responses to singlet oxygen in plants. *Proc. Natl. Acad. Sci. USA*, **109**, 5535–5540.

Ricroch, A.E., Bergé, J.B. and Kuntz, M. (2011) Evaluation of genetically engineered crops using transcriptomic, proteomic, and metabolomic profiling techniques. *Plant Physiol.* **155**, 1752–1761.

Ruhlman, T., Verma, D., Samson, N. and Daniell, H. (2010) The role of heterologous chloroplast sequence elements in transgene integration and expression. *Plant Physiol.* **152**, 2088–2104.

Saxena, B., Subramaniyan, M., Malhotra, D., Bhavesh, N.S., Potlakayala, S.D. and Kumar, S. (2014) Metabolic engineering of chloroplasts for artemisinic acid biosynthesis and impact on plant growth. *J. Biosci.* **39**, 33–41.

Sierro, N., Battey, J.N., Ouadi, S., Bakaher, N., Bovet, L., Willig, A., Goepfert, S. et al. (2014) The tobacco genome sequence and its comparison with those of tomato and potato. *Nat. Commun.* **5**, 3833.

Singh, N.D., Li, M., Lee, S.B., Schnell, D. and Daniell, H. (2008) Arabidopsis Tic40 expression in tobacco chloroplasts results in massive proliferation of the inner envelope membrane and upregulation of associated proteins. *Plant Cell*, **20**, 3405–3417.

Singh, N.D., Kumar, S. and Daniell, H. (2015) Expression of β -glucosidase increases trichome density and artemisinin content in transgenic *Artemisia annua* plants. *Plant Biotechnol. J.* **14**, 1034–1045.

Tattersall, D.B., Bak, S., Jones, P.R., Olsen, C.E., Nielsen, J.K., Hansen, M.L., Høj, P.B. et al. (2001) Resistance to an herbivore through engineered cyanogenic glucoside synthesis. *Science*, **293**, 1826–1828.

Trapnell, C., Pachter, L. and Salzberg, S.L. (2009) TopHat: discovering splice junctions with RNA-seq. *Bioinformatics*, **25**, 1105–1111.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L. et al. (2010) Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515.

Verma, D. and Daniell, H. (2007) Chloroplast vector systems for biotechnology applications. *Plant Physiol.* **145**, 1129–1143.

Verma, D., Samson, N.P., Koya, V. and Daniell, H. (2008) A protocol for expression of foreign genes in chloroplasts. *Nat. Protoc.* **3**, 739–758.

Viitanen, P.V., Devine, A.L., Khan, M.S., Deuel, D.L., Ven Dyk, D.E. and Daniell, H. (2004) Metabolic engineering of the chloroplast genome using the *Escherichia coli* ubiC gene reveals that chorismate is a readily abundant plant precursor for p-hydroxybenzoic acid biosynthesis. *Plant Physiol.* **136**, 4048–4060.

Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63.

Wang, X., Ort, D. and Yuan, J.S. (2015) Photosynthetic terpene hydrocarbon production for fuels and chemicals. *Plant Biotechnol. J.* **13**, 137–146.

Woodson, J.D., Perez-Ruiz, J.M. and Chory, J. (2011) Heme synthesis by plastid ferrochelatase I regulates nuclear gene expression in plants. *Curr. Biol.* **21**, 897–903.

Wu, S., Schalk, M., Clark, A., Miles, R.B., Coates, R. and Chappell, J. (2006) Redirection of cytosolic or plastidic isoprenoid precursors elevates terpene production in plants. *Nat. Biotechnol.* **24**, 1441–1447.

Wu, S., Jiang, Z., Kempinski, C., Nybo, S.E., Husodo, S., Williams, R. and Chappell, J. (2012) Engineering triterpene metabolism in tobacco. *Planta*, **236**, 867–877.

Xiao, Y., Savchenko, T., Baidoo, E.E., Chehab, W.E., Hayden, D.M., Tolstikov, V., Corwin, J.A. et al. (2012) Retrograde signaling by the plastidial metabolite MECPP regulates expression of nuclear stress-response genes. *Cell*, **149**, 1525–1535.

Ye, X., Al-Babili, S., Klöti, A., Zhang, J., Lucca, P., Beyer, P. and Potrykus, I. (2000) Engineering the provitamin A (β -carotene) biosynthetic pathway into (carotenoid-free) rice endosperm. *Science*, **287**, 303–305.

Supporting information

Additional Supporting information may be found in the online version of this article:

Figure S1 PCR analysis of transgene integration in transplastomic plants.

Figure S2 GO analysis for ‘cellular component’-related terms.

Figure S3 GO analysis for ‘molecular function’-related terms.

Figure S4 Plot of \log_2 fold changes for selected transcripts as determined by qRT-PCR versus RNA-seq.

Table S1 Metabolites with significantly changed abundance in each line relative to WT.

Table S2 List of all metabolites whose abundance changes significantly and by at least twofold in PH-SQS compared with PH.

Table S3 List of qRT-PCR primers used in this study.